Development and Evaluation of Search Tasks for IIR Experiments using a Cognitive Complexity Framework

Diane Kelly, Jaime Arguello, Ashlee Edwards and Wan-ching Wu
School of Information and Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC, USA
[dianek, jarguell, aedwards, wanchinw] @email.unc.edu

ABSTRACT

One of the most challenging aspects of designing interactive information retrieval (IIR) experiments with users is the development of search tasks. We describe an evaluation of 20 search tasks that were designed for use in IIR experiments and developed using a cognitive complexity framework from educational theory. The search tasks represent five levels of cognitive complexity and four topical domains. The tasks were evaluated in the context of a laboratory IIR experiment with 48 participants. Behavioral and self-report data were used to characterize and understand differences among tasks. Results showed more cognitively complex tasks required significantly more search activity from participants (e.g., more queries, clicks, and time to complete). However, participants did not evaluate more cognitively complex tasks as more difficult and were equally satisfied with their performances across tasks. Our work makes four contributions: (1) it adds to what is known about the relationship among task, search behaviors and user experience; (2) it presents a framework for task creation and evaluation; (3) it provides tasks and questionnaires that can be reused by others and (4) it raises questions about findings and assumptions of many recent studies that only use *behavioral signals* from search logs as evidence for task difficulty and searcher satisfaction, as many of our results directly contradict these findings.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval: Search Process

Keywords

Search tasks, user studies, interactive IR, search behavior

1. INTRODUCTION

Search tasks are one of the most important components of information search studies. In most experimental studies, researchers assign search tasks to people in order to study search behavior and evaluate systems. In some cases, search tasks are ancillary to the study purposes but are needed for study participants to exercise systems, while in other cases search tasks act as independent variables. Despite their importance, there is little formal guidance about how to construct and evaluate search tasks and empirical reports often do not provide thorough descriptions of how search tasks were generated or full descriptions of the tasks, which inhibits reuse. Moreover, the lack of parity in search tasks across studies generates incommensurable results, which ultimately makes it difficult to clearly understand and generalize task and system effects and replicate results.

The development of search tasks can be difficult and time consuming, and often requires specialized knowledge and skills. Search task development is further complicated by the abundance of research demonstrating how variations in search tasks and search task properties can impact searcher behavior [36, 42]. Poor task design can confound results, generate undesirable search behaviors, create additional variables that complicate analysis, and ultimately, result in wasted time and money. Consider a researcher who is interested in evaluating a search interface designed to support exploratory search. Despite a good faith effort to create search tasks that require sustained interaction, the researcher might inadvertently create tasks that can be addressed with a single, easily findable document such as a Wikipedia page. Thus, even the most well designed experiment might be sabotaged by inappropriately designed search tasks.

Task-based search has been identified as a key research direction at several recent meetings, including the Second Strategic Workshop on Information Retrieval (SWIRL) [2]. There have also been long-standing calls for the development of standardized task sets, reference tasks and sharable tasks that can be used in information search studies [23, 36, 41]. At a recent workshop focused on task-based information search systems, the development of simulated search tasks that could be shared among research groups was identified as a major direction [24]. Specifically, the report states that a re-usable set of search tasks and questionnaires would help make user studies more reproducible and allow for future meta-analysis. Similar recommendations regarding sharable materials for IIR studies were made at an earlier NII Shonan Meeting [9], which focused on whole-session evaluation, as well as an earlier workshop on information-seeking support systems [25].

In this paper, we describe an evaluation of 20 search tasks that were designed for use in IIR experiments and developed using a cognitive complexity framework from educational theory. The hope in using this framework was that we would be able to develop a set of search tasks that would induce a range of varied search behaviors in participants. We use Borlund's [10] notion of simulated work tasks, which specifies that tasks be tailored to target participants, who were undergraduate students in this study. This is one of the most common types of participants in IIR experiments and a population that often performs Internet searches. Behavioral and self-report data were collected to understand and characterize differences among tasks, including with respect to difficulty, engagement and satisfaction.

ICTIR '15, September 27 - 30, 2015, Northampton, MA, USA

ACM 978-1-4503-3833-2/15/09…$15.00
DOI: http://dx.doi.org/10.1145/2808194.2809465

2. BACKGROUND

Several researchers have contributed work that enhances our abilities to discuss, define, create and observe tasks. In one of the first reviews of task-based information seeking research, Vakkari [38] defines a task as an "activity to be performed in order to accomplish a goal" (p. 416). Toms [36] defines a task as having "a defined objective or goal with an intended and potentially unknown outcome or result, and may have known conditional and unconditional requirements" (p. 45). Both Vakkari's and Toms's definitions go beyond an individual search task and focus instead on the larger goals of the user. Byström and Hansen [13] distinguish between a work task and a search task, one or more of which might be conducted to address the work task. Researchers have also articulated definitions for search tasks. Wildemuth, et al. [42] state: "search tasks are goal-directed activities carried out using search systems" (p. 1134). Li and Belkin [27] define an information search task as "a task that users need to accomplish through effective interaction with information systems" (p. 1823). Both of these definitions restrict search tasks to activities done with information systems.

Tasks have been classified according to type (e.g., open, factual, navigational, decision-making) and according to properties (e.g., difficulty, urgency, structure, stage). Toms [36] observes that tasks have been used in two major ways in IR research: (1) as a vehicle for research (citing, for example TREC topics) and (2) as an object of study, where the researcher is interested in how different task types or properties impact search experiences. Li and Belkin [27] unify a variety of task characteristics in a faceted classification of tasks. This classification includes *generic* facets of tasks (e.g., source of task, time, product, process and goal) and *common* facets of tasks including characteristics (e.g., objective task complexity and interdependence) as well as users' perception of task (e.g., salience, urgency, difficulty, subjective task complexity and knowledge of task topic).

Broder's [12] task classification was one of the first in the context of Web search and resulted in three categories of tasks: navigational, informational and transactional. While this classification provided some insight into what people were doing on the Web at the time, its coarseness makes it less useful for designing tasks for IIR studies since so many different types of tasks are grouped together in the informational class. Tasks in this class were also characterized as having single source solutions ("I want a good site on this topic" p. 6). This work launched an interest in navigation and other fact-finding types of tasks, especially since these seemed to be the most common types of tasks people were conducting on the Web. Later, researchers' interests expanded to include more open tasks, such as exploratory tasks. White and Roth [40] provide a discussion of these types of tasks and situate them in the context of other task types.

One of the most common approaches to studying tasks is by separating tasks into types (e.g., fact-finding vs. information gathering; known-item vs. exploratory). For example, Jiang, et al. [22] studied four types of tasks that were created using Li and Belkin's [27] classification: known item, known subject, interpretive and exploratory. They found participants were more active for known subject and interpretive tasks, but issued more queries for known item and exploratory tasks. Toms, et al. [37] examined decision, fact-finding and information gathering tasks, as well as tasks with varying structure: parallel or hierarchical.

Over the years, those creating test collections have also defined and examined a number of different types of tasks, although terminology sometimes differs. TREC has a notion of a topic, which assumes most requests are topical or subject-based [33] and task is used to describe the specific TREC Track focus (e.g., filtering, clinical decision support, microblog). Those studying IIR have used these collections in studies with research participants, although not without some challenges [23]. The TREC Interactive Track also generated and investigated a variety of task types including ad-hoc, filtering, aspectual recall and fact-finding [18]. In 2015, a TREC Track focused on inferring tasks from search behavior[1] has been created.

Another common way that tasks have been characterized in IIR studies is according to difficulty and complexity. Wildemuth, et al. [42] analyze how researchers have conceptually and operationally defined task complexity and task difficulty in a review of over one-hundred IIR studies. They note that despite widespread usage of these concepts, "clear and consistent definitions of these attributes are lacking and there is no consensus on how to distinguish levels of complexity or difficulty within a set of search tasks" (p. 1119). When creating search tasks, each of these attributes needs to be considered separately, and one needs to make a determination about whether one plans to manipulate these attributes or hold them constant. Measurement and manipulation rests on having a solid definition of each construct and a method to consistently measure and observe variations. Wildemuth, et al. go on to note a lack of transparency in how task attributes are measured; task difficulty, in particular, is often conceptualized as a subjective attribute that describes the user's experiences conducting the task and is measured with questionnaires. The content of such questionnaires is often omitted from published reports. Task difficulty is also often measured with a single item; this gross signal lacks precision and likely conflates many different components of difficulty.

One of the more popular conceptualizations of task complexity in the IIR literature comes from the work of Campbell [15], who considered complexity as an objective task characteristic that could be described by four dimensions: (1) the number of potential paths to the desired outcome; (2) the presence of multiple desired outcomes; (3) the presence of conflicting interdependencies between paths; and (4) uncertainty regarding paths. These dimensions are interesting to consider in the context of search task complexity as they suggest that tasks, which have more than one possible solution and allow people to arrive at solutions in many different ways, are more complex. Moreover, the idea of interdependencies between paths suggests that as searchers progress through a task they make a sequence of interconnected decisions that increasingly funnels their focus. Searchers cannot be certain the paths they select will lead to successful solutions until they arrive at the end or may need to shift between paths to resolve these interdependencies.

Another popular conceptualization of task complexity comes from Byström and Järvelin [14] who define task complexity as the *a priori determinability* of tasks, which is the extent to which the searcher can deduce the required task inputs, processes, and outcomes based on the initial task statement. Importantly, this conceptualization was in the context of work tasks, not search tasks. Vakkari [38] described complexity as "the degree of predeterminability of task performance" (p. 826). These conceptualizations are based on subjective task complexity since people make these determinations before conducing a task; Wildemuth, et al. [42] found fewer studies treated search task complexity as a subjective construct, but there is some evidence

[1] http://www.cs.ucl.ac.uk/tasks-track-2015/guidelines.html

that objective and subjective complexity are related. For example, Bell and Ruthven [8] used Byström and Järvelin's conceptualization to create artificial search tasks with different levels of task complexity and found that objective task complexity as they had manipulated it, was correlated with users' subjective assessments of complexity.

Wildemuth, et al. [42, p. 1112] identified six ways that task complexity has been defined and operationalized in the IIR literature: (1) number of subtasks or steps required; (2) number of subtopics or facets; (3) number of query terms and operations required; (4) number of sources or items required; (5) the indeterminate nature of the task; and (6) the cognitive complexity of the task. Examples of work using each of these operationalizations can be found in Wildemuth, et al. In general, studies have found that people engage in more search interactions when completing more complex tasks [28, 34]. For example, Li and Belkin [28] found that objective task complexity was related to number of queries issued, mean query length, number of pages viewed, and number of sources consulted.

Jansen et al. [21] used Anderson and Krathwohl's taxonomy of educational objectives [3] to create complex search tasks reflecting six types of cognitive processes: *remember*, *understand*, *apply*, *analyze*, *evaluate* and *create* (see Table 1 for definitions). Jansen et al. observed a number of significant differences in the amount of interaction users exhibited when completing different task types, including session duration, number of queries, and number of pages viewed. However, the distinctions among the tasks were not clear and increases in cognitive complexity did not always result in increased search behavior. Most notably, participants who completed tasks requiring the highest level of cognitive processing only spent about four minutes completing these tasks. Finally, although compelling, it was not obvious how this taxonomy might be used to create new tasks or how one might use Jansen et al.'s general task structure to create new tasks.

Tasks have also been characterized according to difficulty. Wildemuth, et al. [42] noted few studies that examine both task complexity and task difficulty and some confusion in the literature about terminology (e.g., one person's complexity is another person's difficulty). Campbell [15] conceptualized task difficulty as a subjective characteristic that is a result of the person and task interaction. Wildemuth, et al. make the same distinction and extend the definition of difficulty by describing it as a "relationship between the search task and the searcher or between the search task and the corpus being searched that express the amount of effort of skill required and the likelihood of success" (p.1134). Wildemuth, et al. [p. 1129] identified four major approaches to measuring task difficulty in the literature: (1) as a function of searcher performance; (2) the match between terms in the task description and in the target page; (3) the number of relevant documents in the collection; and (4) the searchers' or experts' perceptions of difficulty, which is arguably the most common way to measure task difficulty.

Papers exploring task difficulty have claimed a positive correlation between task difficulty and interaction [19, 26, 29]. For example, Liu, et al. [29] found that when completing more difficult tasks, participants took longer, entered more queries, viewed more pages, and used more sources. In a follow-up study, Liu, et al. [30] asked their participants to describe what made a search task difficult. Many of the reasons given by participants were based on uncertainty at the outset of the task and problems completing the task. For example, participants gave reasons such as uncertainty about how much effort would be involved with searching for the information, what to do to perform the search, and difficulty formulating queries.

Although not included in Wildemuth, et al.'s [42] review, there are several studies that have used tasks that are insurmountable, or nearly insurmountable; that is, the researchers knew the tasks could not be solved, or solved easily [1, 6]. This is slightly different from selecting tasks based on the number of relevant documents in the collection because both of these studies were done in the context of the Web where the collection was not as defined (most of Wildemuth, et al.'s examples for this approach used TREC test collections). Another interesting case is Smith [36] who used multiple ambiguous terms in the task descriptions to make them difficult. However, all of these cases could still be described by Wildemuth, et al.'s definition of task difficulty.

3. SEARCH TASK DESIGN

To create search tasks, we started from Anderson and Krathwohl's taxonomy of educational objectives [3]. This taxonomy has dimensions to reflect cognitive process and knowledge. Like Jansen et al. [21], we focus on the *cognitive process* dimension (Table 1). There are six types of cognitive processes: *remember*, *understand*, *apply*, *analyze*, *evaluate* and *create*. Each requires increasing amounts of cognitive effort. While this taxonomy is traditionally used to create educational materials such as exercises, we used it as a framework to construct search tasks.

Table 1. Anderson and Krathwohl's Taxonomy of Learning Objectives (Cognitive Process Dimension).

Process	Definition
Remember	Retrieving, recognizing, and recalling relevant knowledge from long-term memory.
Understand	Constructing meaning from oral, written, and graphic messages through interpreting, exemplifying, classifying, summarizing, inferring, comparing, and explaining.
Apply	Carrying out or using a procedure through executing or implementing.
Analyze	Breaking material into constituent parts, determining how the parts relate to one another and to an overall structure or purpose through differentiating, organizing, and attributing.
Evaluate	Making judgments based on criteria and standards through checking and critiquing.
Create	Putting elements together to form a coherent or functional whole; reorganizing elements into a new pattern or structure through generating, planning, or producing.

We selected four domains to use when creating the tasks: health, commerce, entertainment, and science and technology, and following Borlund [10], situated the tasks within scenarios geared toward our target participants, university undergraduates. In many cases, we selected topics that were of regional interest and used informal language when constructing scenarios. We created 20 tasks: one for each cognitive domain and cognitive process, except *apply* because we were unable to create search tasks for this category that were distinct from the other categories. The tasks are available online at http://ils.unc.edu/searchtasksforiir/. Examples from two domains are shown in Table 2.

The tasks differ on two dimensions: the type of information desired (target outcome) and the types of activities that need to be completed. The cognitive processes build on one another as in [3] (see Table 2). *Remember* tasks require a specific fact for their resolution, such as a number or location. These types of tasks

require the searcher to identify or recognize the fact as it occurs in an information source. *Understand* tasks require the searcher to provide an exhaustive list of items; for example, health risks. Similar to remember tasks, this type of task primarily requires the searcher to identify a list or factors in an information source and possibly compile the list from multiple sources if a single list cannot be found. *Analyze* tasks require the searcher to find and compile a list of items and to understand and describe their differences. *Evaluate* tasks require the searcher to find and compile a list of items, understand their differences and make a recommendation. The target outcome for *create* tasks is a plan, which requires the searcher to perform the same sets of actions for the evaluate tasks, except instead of a justification the searcher needs to generate something.

Table 2. Cognitive Processes, Target Outcomes, Mental Activities and Example Tasks

Process	Target Outcome	Mental Activities
Remember (R)	Fact	Identify
You recently watched a documentary about people living with HIV in the United States. You thought the disease was nearly eradicated, and are now curious to know more about the prevalence of the disease. Specifically, how many people in the US are currently living with HIV?		
Understand (U)	List (set)	Identify, Compile
Your nephew is considering trying out for a football team. Most of your relatives are supportive of the idea, but you think the sport is dangerous and are worried about the potential health risks. Specifically, what are some long-term health risks faced by football players?		
Analyze (A)	List (prioritized), Description	Identify, Compile, Describe
Having heard some of the recent reports on risks of natural tanning, it seems like a better idea to sport an artificial tan this summer. What are some of the different types of artificial tanning methods? What are the health risks associated with each method?		
Evaluate (E)	Recommendation	Identify, Compile, Describe, Compare, Decide, Justify
One of your siblings got a spur of the moment tattoo and now regrets it. What are the current available methods for tattoo removal, and how effective are they? Which method do you think is best? Why?		
Create (C)	Plan	Identify, Compile, Describe, Compare, Decide, Make
After the NASCAR season opened this year, your niece became really interested in soapbox derby racing. Since her parents are both really busy, you've agreed to help her build a car so that she can enter a local race. The first step is to figure out how to build a car. Identify some basic designs that you might use and create a basic plan for constructing the car.		

These tasks can be related to other conceptualizations of task complexity in the information search literature. Using Campbell's [15] conceptualization of objective task complexity, *Remember* tasks have the fewest solution paths, the fewest number of solutions or outcomes, the least amount of conflicting interdependencies between paths and the least amount of uncertainty regarding paths, while *create* tasks have the greatest numbers. The tasks can also be related to Byström and Järvelin's [14] definition of task complexity in that the expected inputs, processes and outcomes become less certain as one moves from *Remember* to *Create* tasks. Finally, our target outcomes are similar to Li and Belkin's [27] product facet. One of the unique aspects of our framework is that it takes the abstract features described by other researchers and presents them more concretely by providing task descriptions that can be more easily reused, either by using the tasks in their current forms, or by extrapolating templates with slots that can be filled in with concepts that are tailored to target participants.

Several of the tasks from this study were initially created and used in another research project [5]. The initial development consisted of pilot tests with six participants and then a full user study with 28 participants from the local community with tasks representing the first three levels of cognitive complexity. In this previous study, a significant relationship was found between cognitive complexity and search behaviors: as complexity increased, so did the number of queries issued, URLs visited, search results clicked and time spent conducting the search. The tasks were updated and revised based on the findings from this study and underwent additional critique and analysis. New versions of the tasks were then used in the current study. Preliminary results of this current study have been presented in a poster paper [43]; however, the poster paper described a subset of measures from the first twenty-four participants and for half of the search tasks.

4. METHOD

A laboratory study was conducted to evaluate the search tasks. Each participant completed five search tasks (one representing each cognitive complexity level) from a single domain. Tasks were rotated using a Latin-square. Participants searched the open Web using the search engine of their choice (in all cases, Google) and were asked to create responses to each task by typing answers and/or copying and pasting evidence that helped them arrive at their answers. No task time limits were imposed. Participants were given a monetary honorarium at the end of the study session.

4.1 Pre-Search Questionnaire

Participants completed a pre-search questionnaire before each search with items about interests and knowledge, task complexity and expected task difficulty (Table 3). The scale anchors for the interests and knowledge items can be viewed in Table 6. Unless otherwise specified, all other items were evaluated with a five-point scale, where 1=not at all, 2=slightly, 3=somewhat, 4=moderately and 5=very. Three items were included to measure participants' perceptions of task complexity; these items were based on Byström and Järvelin's [14] definition of task complexity. Although we consider task complexity to be an objective property inherent to tasks we included these items to see how participants would evaluate these attributes.

Expected task difficulty was related to participants' expectations about the potential challenges associated with completing search activities, including results evaluation and determining when they had enough information to stop searching. We use these same items in the Post-Task Questionnaire to measure if and how participants' assessments changed after completing the search tasks. While task complexity relates to the task description, expected difficulty relates to the participant's expected search experience. Thus, in this study, we manipulate cognitive complexity through task design and measure task complexity and task difficulty, via pre- and post-search questionnaires. This allowed us to see how all of these measures were related to each other as well as to search behavior.

Table 3. Pre-Task Questionnaire Items

Interest & Knowledge	How interested are you to learn more about the topic of this task?
	How many times have you searched for information about this task?
	How much do you know about the topic of the task?
Task Complexity	How defined is this task in terms of the types of information needed to complete it?
	How defined is this task in terms of the steps required to complete it?
	How defined is this task in terms of its expected solution?
Expected Task Difficulty	How difficult do you think it will be to *search* for information for this task using a search engine?
	How difficult do you think it will be to *understand* the information the search engine finds?
	How difficult do you think it will be to *decide* if the information the search engine finds is *useful* for completing the task?
	How difficult do you think it will be to *integrate* the information the search engine finds?
	How difficult do you think it will be to determine *when you have enough* information to finish the task?

4.2 Post-Search Questionnaire

The Post-Task Questionnaire was divided into five parts (Table 4). The first part asked participants to describe how they felt while completing the task. The second part asked participants to indicate the extent to which their interests and knowledge changed. The third part consisted of the five difficulty items from the Pre-Search Questionnaire with minor editorial changes to reflect past tense. The fourth and fifth parts elicited summative judgments from participants about difficulty and satisfaction.

Table 4. Post-Task Questionnaire Items

Engagement	How enjoyable was it to do this task?
	How engaging did you find this task?
	How difficult was it to concentrate while you were doing this task?
Interest	How much did your interest in the task increase as you searched?
	How much did your knowledge of the task increase as you searched?
Experienced Task Difficulty	Same five items from Table 3 except items started with, "How difficult was it to …"
Overall Difficulty	Overall, how difficult was this task?
Overall Satisfaction	Overall, how satisfied are you with your solution to this task?
	Overall, how satisfied are you with the search strategy you took to solve this task?

4.3 Exit Questionnaire

The Exit Questionnaire asked participants to rank the tasks according to difficulty and engagement (1=least; 5=most) and to explain their rankings.

4.4 Search Behaviors

Participants' searches were logged using the Lemur Query Toolbar and 11 measures were computed from this log (Table 5). Measures 1-8 were computed at the search session level; these values were then averaged and are reported according to cognitive complexity level. The last three measures illustrate the extent to which participants who completed a task (complexity-domain combination) deviated from *other* participants who completed the *same* task for queries issued, query-terms used, and URLs visited.

Table 5. Search Behavior Measures

Measure	Definition
Queries	Total number of unique queries submitted by a participant when completing a task.
Query length	Average number of query terms in all unique queries issued for a task.
Unique query terms	Total number of unique query terms used by a participant when completing a task.
SERP clicks	Total number of clicks participants made on SERPs.
URLs visited	Total number of unique URLs visited by participants (includes URLs accessed directly and indirectly via SERP)
Queries w/o SERP clicks	Total number of unique queries where participants did not click on the search engine results page (SERP).
Time to completion	The amount of time (in seconds) participants spent completing search tasks.
SERP dwell time	Average time spent between issuing a new query and clicking on the first search result (in seconds).
Query diversity	Number of queries issued that were not issued by another participant completing the exact same task.
Query term diversity	Number of query terms used that were not used by another participant completing the exact same task.
URL diversity	Number of URLs visited that were not visited by another participant completing the exact same task.

4.5 Participants

Forty-eight undergraduate participants were recruited from our university via mass email solicitation. Thirty-three participants were female and 15 were male. Participants' average age was 20 years old (SD=1.62). The frequency of majors was 10 sciences, 28 social sciences, 3 humanities, 6 professional schools and 1 undecided. Most participants reported conducting information searches daily with an average of 7-9 years of search experience.

4.6 Data Analysis

Unless otherwise specified, repeated-measures ANOVAs were conducted with cognitive complexity level as a within-subject factor. Bonferroni tests were used as the follow-up tests. Alpha was set to 0.01 for all analyses.

5. RESULTS

5.1 Search Behaviors

Figure 1 displays participants' mean search behaviors according to cognitive complexity level, statistical test results (F-tests and follow-up tests) and effect sizes (η^2)[2].

Participants submitted the most queries when completing *create* tasks (M=4.85; SD=4.42) and the fewest when completing *remember* tasks (M=1.68; SD=1.04). Statistical tests showed participants entered significantly fewer queries for *remember* and *understand* tasks than for all other tasks. Significant differences

[2] Search interaction data from one participant was not captured because of technical difficulties and the pre-search questionnaires from another participant were not recorded properly, so analyses of search behaviors and pre-search questionnaire responses are based on 47 participants.

were also detected between *analyze* tasks and *create* tasks, and *evaluate* tasks and *create* tasks.

While participants submitted the most queries for *create* tasks, these queries were on average shorter than the queries they submitted for other tasks (M=4.04; SD=1.37). Participants submitted the longest queries for *remember* tasks (M=6.01; SD=2.94). Statistical tests showed participants entered significantly longer queries for the *remember* tasks than for *analyze* and *create* tasks. Queries submitted when completing *understand* and *evaluate* tasks were also significantly longer than those submitted for *create* tasks.

Participants used the greatest number of unique terms in their queries when completing *create* tasks (M=10.68; SD=6.69) and the least for *remember* tasks (M=7.02; SD=3.12). Significant differences were also detected here, with participants submitting significantly more unique query terms when completing *create* tasks than *remember* or *understand* tasks.

Participants made the most SERP clicks when completing *create* tasks (M=5.98; SD=5.02) and the fewest when completing *remember* tasks (M=2.49; SD=1.56). The general trend was that SERP clicks increased as task complexity increased. A statistically significant relationship was detected between task complexity and SERP clicks; follow-up tests detected reliable differences between *remember* and *understand* tasks and *create* tasks. This relationship was even more pronounced for number of URLs visited, with participants viewing an average of 14.43 (SD=12.34) URLs when completing *create* tasks and 3.70 (SD=3.93) when completing *remember* tasks. With respect to URLs viewed, statistical tests showed participants visited significantly more URLs for *create* tasks than any of the other tasks. Participants also visited significantly more URLs for *analyze* and *evaluate* tasks than for *remember* tasks.

As with the previous measures, there was also a general trend for number of queries without clicks to increase with complexity, although *evaluate* tasks had slightly fewer queries without clicks than *analyze* tasks. Statistical tests showed participants issued significantly more queries without clicks for *create* tasks than *remember* or *understand* tasks.

With respect to query diversity, participants submitted the greatest number of unique queries for *create* tasks (M=4.26; SD=3.90) and the fewest for *remember* tasks (M=1.04; SD=1.00). A significant difference was detected with the differences between *remember* and *understand* tasks and *analyze* and *evaluate* tasks being significant, as well as the differences between *analyze* and *evaluate* tasks and *create* tasks.

When we consider query term diversity, we see that the number of unique terms entered by participants was the greatest for *create* tasks (M=2.40; SD=2.79) and the least for *remember* tasks (M=0.77; SD=1.13), indicating much greater overlap in the terms participants used when completing *remember* tasks. Tests show the query terms used for *analyze* and *create* tasks were significantly more diverse than those used for *remember* tasks.

When we consider URL diversity, we see increasing means as we move from *remember* (M=1.43; SD=3.66) to *create* (M=10.43; SD=10.64). Results showed significant differences in the number of unique URLs visited by participants between all pairs of task except *analyze* and *evaluate* tasks.

Our diversity measures considered the *absolute* number of queries, query-terms and URLs that were not observed in another search session for the same task (complexity-domain combination). Given that task complexity had a significant effect on the number of queries and query terms used, and URLs visited, we also computed normalized versions of our diversity measures (not shown in Figure 1). The normalized versions considered the *percentage* of queries, query-terms, and URLs that were not observed in another session for the same task. ANOVAs using these normalized values were also statistically significant.

The time participants spent completing tasks increased with task complexity, with participants spending the most time completing *create* tasks (M=9.868m; SD=5.295m) and least time completing *remember* tasks (M=2.838m; SD=2.453m). Statistical tests showed participants spent significantly more time completing *analyze*, *evaluate* and *create* tasks than *remember* or *understand* tasks, and significantly more time completing *understand* than *remember* tasks. With respect to mean SERP dwell time, there did not appear to be any trend. For most tasks, participants spent around 8 seconds viewing SERPs with the exception being *remember* tasks where they spent about 5.3 seconds; no significant differences were detected.

5.2 Task Knowledge & Interest

Table 6 displays the distribution of responses to the pre-task questionnaire items about prior knowledge. For most tasks (78%) participants indicated they had never searched for information about the task. For about 50% of the tasks, participants indicated they knew nothing. Friedman's test revealed no significant differences in these distributions.

Table 6. Frequency of responses for prior task knowledge.

Times Searched, Knowledge	Never, Nothing	1-2 times, Little	3-4 times, Some	5+ times, Great deal
Remember	81%, 64%	15%, 21%	<1%, 13%	<1%, <1%
Understand	81%, 38%	15%, 44%	<1%, 17%	0%, 0%
Analyze	79%, 43%	19%, 47%	<1%, 1%	0%, <1%
Evaluate	72%, 38%	21%, 34%	1%, 23%	0%, <1%
Create	79%, 60%	1%, 28%	1%, 10%	<1%, <1%

Participants were asked in the post-search questionnaire about the extent to which their knowledge of the tasks increased after searching (Table 7). For most tasks, participants indicated their knowledge increased somewhat. No significant differences were found in participants' responses to this item according to complexity level. Participants' interests in the tasks were elicited in the pre- and post-search questionnaires (Table 7). While there were no significant differences in their pre-search interest ratings, there was a significant difference in post-search ratings [$F(4, 188)=4.09$, $p=0.003$, $\eta^2=0.08$]: participants were significantly more interested in the *evaluate* and *create* tasks than the *remember* tasks. A comparison of the pre- and post-search ratings also shows that interest levels decreased for *remember* and *understand* tasks, but increased for *create* tasks.

Table 7. Mean (SD) Interest and Knowledge Increase. *p<0.01

	Remember	Understand	Analyze	Evaluate	Create
Pre-Interest	2.70 (1.25)	2.94 (1.17)	2.70 (1.12)	2.91 (1.30)	2.45 (1.35)
Post-Interest*	2.23 (1.36)	2.54 (1.17)	2.77 (1.15)	2.94 (1.13)	3.02 (1.08)
Knowledge Increase	2.92 (1.18)	3.21 (1.24)	3.15 (1.13)	3.27 (1.14)	3.29 (1.07)

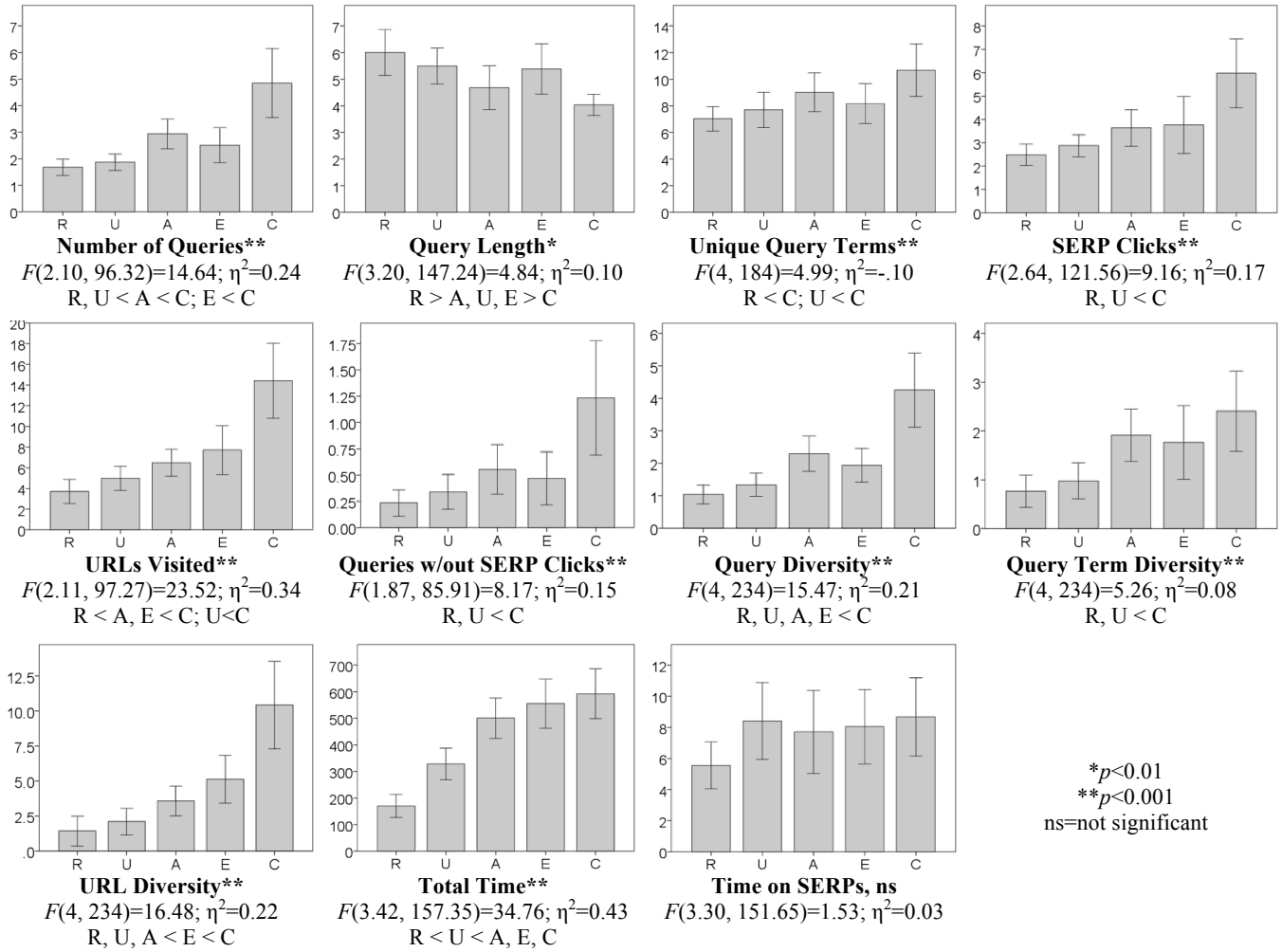
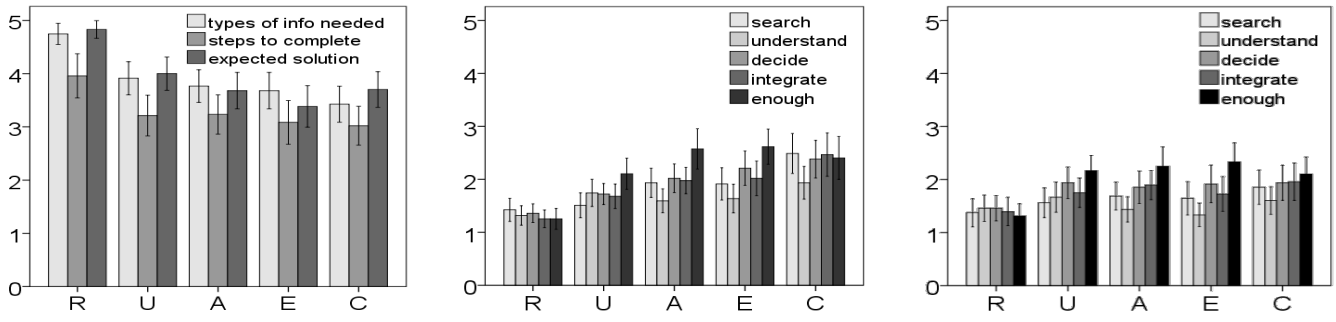


Figure 1. Search behaviors (means, 95% confidence intervals) according to cognitive complexity level. Varying degrees of freedom exist in cases where Mauchly's test of sphericity indicated different variances.



Pre-Search Task Complexity

Types**: F=13.48, η^2=0.23, R > U, A, E, C
Steps**: F=5.49, η^2=0.11, R > U, A, E, C
Solution**: F=14.46, η^2=0.24, R > U, A, E, C

Pre-Search Task Difficulty

Search**: F=9.72, η^2=0.17, R < A, C; U < C
Understand*: F=3.84, η^2=0.08, R < C
Decide**: F=10.48, η^2=0.19, R < A, E, C; U < C
Integrate**: F=13.38, η^2=0.23, R < U < A < C; R < E
Enough**: F=16.24, η^2=0.26, R < U, A, E, C

Post-Search Task Difficulty

Search: F=1.60, ns
Understand: F=1.22, ns
Decide: F=2.12, ns
Integrate: F=2.57, ns
Enough**: F=8.74, η^2=0.16, R < U, A, E, C

Figure 2. Task complexity and pre- and post-search difficulty ratings (means, 95% confidence intervals) according to cognitive complexity level. Statistically significant differences are noted with *p<0.01 and **p<0.001, ns=not significant. Degrees of freedom for pre-search items (4, 184) and for post-search items (4, 188).

5.3 Task Complexity

Figure 2 shows participants' mean ratings of the task complexity items. In general, as the cognitive complexity increased, so did participants' ratings of complexity (lower ratings mean more uncertainty). Significant differences were found for all items; follow-up tests showed *remember* tasks were rated as significantly more defined than other tasks.

5.4 Task Difficulty

Overall, participants' pre-search difficulty ratings showed they did not expect tasks to be difficult (Figure 2). *Remember* tasks were uniformly rated as easy across all items. For *understand*, *analyze* and *evaluate* tasks, participants expected that it would be more difficult to determine when to stop searching. In general, as the level of task complexity increased, participants anticipated that tasks would be more difficult. Significant effects for task complexity were found for the five difficulty items; specifically, participants anticipated *remember* tasks would be significantly less difficult than *create* tasks. Participants anticipated *remember* tasks would always be the least difficult and *create* tasks the most difficult. Participants' post-search difficulty ratings (Figure 2) exhibited a similar pattern to those of the pre-search difficulty items, however most ratings were lower. Significant differences were only found for the item about determining if one had enough information to stop; participants' ratings of *remember* tasks were significantly lower than their ratings of other tasks.

We examined the extent to which participants' responses to the pre- and post-search task difficulty items differed. For about 50% of tasks, there were no changes to any of the difficulty items. Participants rated about 15% of tasks as more difficult (+1 point) and 6% as much more difficult (+2 points). Participants rated about 20% of tasks as easier (-1 point) and 10% as much easier (-2 points). Paired-sample t-tests showed participants' experienced less search difficulty than they expected [expected: M=1.86, SD=1.04; experienced: M=1.64, SD=1.10, $t(234)$=2.68, p<0.01].

5.5 Enjoyment and Engagement

Overall, participants found all their search tasks somewhat enjoyable (Figure 3). There was a significant difference in participants' engagement ratings [$F(4, 188=5.39$, $p<0.001$, $\eta^2=0.12$]: participants rated the *evaluate* and *create* tasks as significantly more engaging than the *remember* tasks, and the *evaluate* tasks as significantly more engaging than the *understand* tasks. There were no significant differences in participants' abilities to concentrate when completing the different tasks.

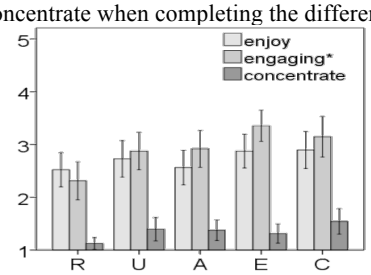


Figure 3. Enjoyment, engagement and concentration. Bars show means and 95% confidence intervals. *p<0.001

5.6 Overall Difficulty and Satisfaction

After rating each task according to the difficulty items, participants provided summative evaluations of task difficulty, satisfaction with solution and satisfaction with search strategy (Figure 4). Participants did not find any of the tasks difficult overall and there were no significant differences according to complexity level. Participants were generally satisfied with their solutions to tasks as well as their strategies.

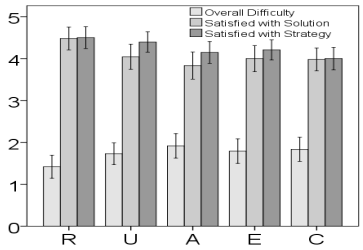


Figure 4. Task difficulty and satisfaction Bars show means and 95% confidence intervals.

5.7 Difficulty and Engagement Rankings

At the end of the study, participants were asked to rank the tasks according to difficulty and engagement (Figure 5). Spearman's Rho showed significant correlations between cognitive complexity and difficulty: ρ=0.413, p<0.0001 and engagement: ρ=0.187, p=0.004, but the effect sizes were small.

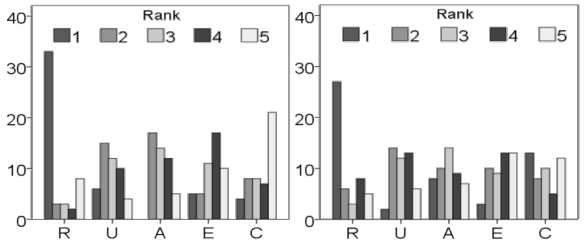


Figure 5. Frequency counts of difficulty (left) and engagement (right) rankings (1=least; 5=most).

The difficulty ranking indicates a fairly clear message about *remember* tasks and while *create* tasks were more often ranked as the most difficult, it is critical to remember that there were no significant differences in participants' ratings of task difficulty. Thus, these rankings should be understood as a relative ordering, which, combined with participants' difficulty ratings, indicate the tasks are clustered in the low area of difficulty. When explaining their difficulty rankings, participants overwhelmingly cited the open-endedness of the task. Many participants described tasks that allowed for more personal input as more engaging. Participants stated that tasks without definite answers were more difficult because this meant they had to make more unique decisions about relevance. While the engagement rankings send a clear message about the *remember* tasks, there was much more variability in the rankings of the other tasks. Participants' explanations of their engagement rankings were overwhelmingly focused on what they found interesting about the task and personal relevance.

6. DISCUSSION & CONCLUSIONS

Results showed when completing tasks of varying levels of cognitive complexity, participants spent significantly different amounts of time and engaged in significantly different amounts of interaction. While there were not always significant differences in search behaviors for tasks of mid-level cognitive complexity (understand, analyze, evaluate), in all cases but one, there were significant differences between the least and most cognitively complex tasks (remember versus create). For most interaction measures, the general trend was for the numbers to increase with cognitively complexity, except for query length, which decreased. An examination of participants' queries for *remember* tasks showed that in many cases participants used most of the information provided in the task description in their queries and sometimes even posed their queries as questions (e.g., What is the

deepest point in the ocean?). Given that *remember* tasks were the most specified, these results are not surprising. Participants submitted the fewest queries for these tasks, which suggests their long queries were useful for addressing these tasks.

Participants submitted the most queries for *create* tasks and used significantly more unique query terms for these tasks than for other tasks. These results suggest participants addressed *create* tasks by submitting a greater number of shorter queries and used more unique terms. These results provide some evidence that the most cognitively complex tasks were similar to Campbell's [15] characterization of complex tasks with respect to having multiple interdependent parts that needed to be addressed separately.

With respect to query diversity, query term diversity, and URL diversity we see even stronger results. As a reminder, these diversity measures compare participants' queries and URLs with those of other participants completing the exact same task. The uniqueness measures discussed above only examined a single participant's behaviors. Queries issued by those completing *create* tasks were significantly more diverse than queries issued by those completing any of the other task types. Participants completing *create* tasks used significantly more diverse terms than when completing *remember* or *understand* tasks, and visited significantly more diverse URLs. These findings provide evidence that the task design encouraged participants to engage in more open-ended, self-directed searching by allowing for more unique and varied solutions. This is consistent with Campbell's [15] description of task complexity.

While more cognitively complex tasks required more time to complete and more interaction (e.g., more SERP clicks and URL visits), they were not associated with higher levels of difficulty or lower levels of satisfaction with the outcome. These results are interesting because they show that the relationship between physical effort and self-reported task difficulty are not linear. These findings question recent work that has proposed such relationships [19, 29]. Other studies have proposed that increases in search behavior are related to dissatisfaction and subsequently search engine switching [39] and seeking help from online QA forums [31]. In our study, more cognitively complex tasks were also associated with more queries without SERP clicks. Query abandonment, or queries without clicks, has also been used as a sign of dissatisfaction, or failure [17]. Our results show the relative nature of interaction signals: in some situations a lot of interaction might indicate problems, while in other cases it represents satisfactory progress or positive experiences, which supports recent work on disambiguating interaction data [20].

In the context of task difficulty, these results lead us to speculate that search task difficulty is primarily a function of what a person expects when starting a task. Specifically, we posit that when searchers encounter a search task description, they appraise the task and its requirements in the context of their abilities, desires and other aspects of the search situation (e.g., system familiarity, time limits) and then estimate how much effort they believe is needed to complete the task at their ideal level of performance. A search task, consequently, becomes "difficult" when expended effort exceeds expected effort. Additional research is needed to evaluate this theory. One important aspect of the search situation in IIR experiments is that tasks are typically assigned to participants, which likely impacts how participants evaluate difficulty and the extent to which they engage in search.

We measured both task complexity and task difficulty and maintained distinct definitions of these concepts as recommended by Wildemuth et al. [42]. To measure complexity, we used three items that reflected Byström and Järvelin's [14] definition and found that overall, participants' rated all the tasks as fairly well-defined (i.e., not complex) and that *remember* tasks were rated as significantly less complex than the other tasks. With respect to measurement of task difficulty, our results also show that difficulty ratings and rankings provide different types of information. While the rankings showed significant differences among the tasks, the ratings did not. These results show none of our tasks were perceived as very difficult, but that participants could still order them relative to one another, and this ordering was likely within the context of the low end of the task difficulty scale. This finding suggests researchers should use caution when asking participants to rank items, since rankings might be misleading. These findings also provoke questions about the range of task complexity and task difficulty that can be observed within laboratory settings.

In this paper, we addressed long-standing calls for the development of shared infrastructure, in particular, search tasks and questionnaires for conducting IIR studies. We presented a set of search tasks, an analytical description of their design (Table 2) and evidence about the types of behaviors these tasks elicit from research participants. We recognize our tasks have some cultural biases built-in; this was necessary in order to create tasks that we thought would appeal to our target participants. We also focused on one type of searcher that is often represented in IIR studies, undergraduate students. Future studies might investigate our tasks with different populations of participants to see if the general relationships hold, especially with regard to search interaction. In our initial work with these tasks, we studied people from our community and found similar results [5]. Since conducting the research reported in this paper, we have used these tasks in other research projects [4, 11, 16]; several of these studies used a crowdsourcing platform and one was done at a public library with members of our local community. The general findings with respect to interaction have been repeated many times, with only slight variations. For example, crowd-sourced participants did not spend as long overall completing the tasks, but the relative differences among cognitive complexity levels remained. Other researchers have also started to use these tasks [7, 32], which demonstrate their potential usefulness for enabling research. We provide our task complexity and task difficulty questionnaires in this paper, which allows for reuse, critique and further development. While we have yet to perform any psychometric analysis of these questionnaire items to establish validity and reliability, we plan to do so in the future.

7. REFERENCES

[1] Ageev, M., Guo, Q, Lagun, D. & Agichtein, E. (2011). Find it if you can: A game for modeling different types of web search success using interaction data. *Proc. of SIGIR*, 345-354.

[2] Allan, J., Croft, B., Moffat, A. & Sanderson, M. (Eds). (2012). Frontiers, challenges and opportunities for Information retrieval: Report from SWIRL 2012. *SIGIR Forum, 46(1),* 2-32.

[3] Anderson, L. W. & Krathwohl, D. A. (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

[4] Arguello, J. (2014). Predicting search task difficulty. *Proc. ECIR,* 88-99.

[5] Arguello, J., Wu, W.C., Kelly, D., & Edwards, A. (2012). Task complexity, vertical display and user interaction in aggregated search. *Proc. of SIGIR,* 435-444.

[6] Aula, A., Khan, R. M. & Guan, Z. (2010). How does search behavior change as search becomes more difficult? *Proc. of SIGCHI Conference*, 35-44.

[7] Bailey, P., Moffat, A., Scholer, F., & Thomas, P. (2015). User Variability and IR System Evaluation. *Proc. of SIGIR*.

[8] Bell, D. J. & Ruthven, I. (2004). Searcher's assessments of task complexity for web searching. *Proc. of ECIR*, 57-71.

[9] Belkin, N. J., Dumais, S. Kando, N. & Sanderson, M. (2012, October). *Whole Session Evaluation of Interactive Information Retrieval Systems*. National Institute of Informatics Shonan Meeting, Shonan Village Center, Japan.

[10] Borlund, P. (2003). The IIR evaluation model: A framework for the evaluation of interactive information retrieval systems. *Information Research, 8*(3), paper 152.

[11] Brennan, K., Kelly, D., & Arguello, J. (2014). The effect of cognitive abilities on information search for tasks of varying levels of complexity. *Proc. of IIiX,* 165-174.

[12] Broder, A. (2002). A taxonomy of web search. *SIGIR Forum, 36*(2), 3-10.

[13] Byström, K. & Hansen, P. (2005). Conceptual framework for tasks in information studies. *JASIST, 56*(10), 1050–1061.

[14] Byström, K. & Järvelin, K. (1995). Task complexity affects information seeking and use. *IP&M*, *31*, 191-213.

[15] Campbell, D. J. (1988). Task complexity: A review and analysis. *Academy of Management Review*, *13*, 40-52.

[16] Crescenzi, A., Capra, R. & Arguello, J. (2013). Time Pressure, User Satisfaction and Task Difficulty. *Proc. of ASIST Conference.*

[17] Diriye, A., White, R. W., Buscher, G., & Dumais, S. T. (2012). Leaving so soon? Understanding and predicting web search abandonment rationales. *Proc. of CIKM*, 1025-1034.

[18] Dumais, S. T., & Belkin, N. J. (2005). The TREC Interactive Tracks: Putting the user into search. In E. M. Voorhees & D. K. Harman (Eds.) *TREC: Experiment and Evaluation in Information Retrieval* (pp. 123-153), Cambridge, MA: MIT Press.

[19] Gwizdka, J. & Spence, I. (2006). What can searching behavior tell us about the difficulty of information tasks? A study of Web navigation. *Proc. of ASIST*, 1-22.

[20] Hassan, A., White, R. W., Dumais, S. T., & Wang, Y. M. (2014). Struggling or exploring?: Disambiguating long search sessions. *Proc. of WSDM*, 53-62.

[21] Jansen, B. J., Booth, D. & Smith, B. (2009). Using the taxonomy of cognitive learning to model online searching. *IP&M*, *45*, 643-663.

[22] Jiang, J., He, D. & Allan, J. (2014). Searching, browsing and clicking in a search session: Changes in user behavior by task and over time. *Proc. of SIGIR*, 607-616.

[23] Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval, 3(1-2)*.

[24] Kelly, D., Arguello, J., & Capra, R. (2013). NSF workshop on task-based information search systems. *SIGIR Forum, 47*(2).

[25] Kelly, D., Dumais, S., & Pedersen, J. (2009). Evaluation challenges and directions for information seeking support systems. *IEEE Computer*, *42*(3), 60-66.

[26] Kim, J. (2006). Task difficulty as a predictor and indicator of web searching interaction. *Proc. of CHI (Extended Abstracts)*, 959-964.

[27] Li, Y. & Belkin, N. J. (2008). A faceted approach to conceptualizing tasks in information seeking. *IP&M*, *44*, 1822-1837.

[28] Li, Y. & Belkin, N. J. (2010). An exploration of the relationship between work task and interactive information search behavior. *JASIST, 61*(9), 1771–1789.

[29] Liu, J., Liu, C., Cole, M., Belkin, N. J., & Zhang, X. (2012). Exploring and predicting search task difficulty. *Proc. of CIKM*, 1313-1322.

[30] Liu, J., Kim, C. S. & Creel, C. (2013). Why do users feel search task difficult? *Proc. of ASIST*.

[31] Liu, Q., Agichtein, E., Dror, G., Maarek, Y. & Szpektor, I. (2012). When web search fails, searchers become askers: Understanding the transition. *Proc. of SIGIR*, 801-810.

[32] Moffat, A., Thomas, P., & Scholer, F. (2013). Users versus models: What observation tells us about effectiveness metrics. *Proc. of CIKM*, 659–668.

[33] Robertson, S. (2008). On the history of evaluation in IR. *Journal of Documentation, 34*(4), 439-456.

[34] Singer, G., Norbisrath, U. & Lewandowski, D. (2012). Ordinary search engine users assessing difficulty, effort, and outcome for simple and complex search tasks. *Proc. of IIiX*, 110-119.

[35] Smith, C.L. (2008). Searcher adaptation: A response to topic difficulty. *Proc. of the ASIST Conference*.

[36] Toms, E. (2011). Task-based information searching and retrieval. In Ruthven, I., & Kelly, D. (Eds.) *Interactive Information-seeking, Behaviour and Retrieval* (pp.43-59).

[37] Toms, E., O'Brien, H. L., MacKenzie, T., Jordan, C., Freund, L., Toze, S., Dawe, E., & MacNutt, A. (2007). Task effects on interactive search: The query factor. *Proc. of INEX,* 359-372.

[38] Vakkari, P. (2003). Task-based information searching. *ARIST, 37*, 413–464.

[39] White, R.W. & Dumais, S. T. (2009). Characterizing and predicting search engine switching behavior. *Proc. of CIKM*, 87-96.

[40] White, R. W. & Roth, R.A. (2009). *Exploratory search: Beyond the query-response paradigm*. Morgan & Claypool.

[41] Whittaker, S., Terveen, L., & Nardi, B. (2000). Let's stop pushing the envelope and start addressing it: A reference task agenda for HCI. *Human Computer Interaction,* 15, 75-106.

[42] Wildemuth, B. W., Freund, L. & Toms, E. G. (2014). Untangling search task complexity and difficulty in the context of interactive information retrieval studies. *Journal of Documentation, 70*(6), 1118-1140.

[43] Wu, W.C., Kelly, D., Edwards, A., & Arguello, J. (2012). Grannies, tanning beds, tattoos and NASCAR: Evaluation of search tasks with varying levels of cognitive complexity. *Proc. of IIiX*, 254-257.